Title
Population structure in genetic studies: Confounding factors and mixed models.

Journal
PLoS genetics, 14(12)

ISSN
1553-7390

Authors
Sul, Jae Hoon
Martin, Lana S
Eskin, Eleazar

Publication Date
2018-12-01

DOI
10.1371/journal.pgen.1007309

Peer reviewed

PLOS GENETICS

REVIEW

Population structure in genetic studies: Confounding factors and mixed models

Jae Hoon Sul[1], **Lana S. Martin**[2], **Eleazar Eskin**[2,3]*

1 Department of Psychiatry and Biobehavioral Sciences, University of California Los Angeles, Los Angeles, California, United States of America, **2** Department of Computer Science, University of California, Los Angeles, California, United States of America, **3** Department of Human Genetics, University of California Los Angeles, Los Angeles, California, United States of America

* eeskin@cs.ucla.edu

Abstract

A genome-wide association study (GWAS) seeks to identify genetic variants that contribute to the development and progression of a specific disease. Over the past 10 years, new approaches using mixed models have emerged to mitigate the deleterious effects of population structure and relatedness in association studies. However, developing GWAS techniques to accurately test for association while correcting for population structure is a computational and statistical challenge. Using laboratory mouse strains as an example, our review characterizes the problem of population structure in association studies and describes how it can cause false positive associations. We then motivate mixed models in the context of unmodeled factors.

Citation: Sul JH, Martin LS, Eskin E (2018) Population structure in genetic studies: Confounding factors and mixed models. PLoS Genet 14(12): e1007309. https://doi.org/10.1371/journal.pgen.1007309

Editor: Gregory S. Barsh, Stanford University School of Medicine, UNITED STATES

Published: December 27, 2018

Funding: E.E. is supported by National Science Foundation grants 0513612, 0731455, 0729049, 0916676, 1065276, 1302448, 1320589, 1331176, and 1815624, and National Institutes of Health grants K25-HL080079, U01-DA024417, P01-HL30568, P01-HL28481, R01-GM083198, R01-ES021801, R01-MH101782 and R01-ES022282. The funders had no role in the preparation of the article.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Genetics studies have identified thousands of variants implicated in dozens of common human diseases [1–5]. These variants are locations in the human genome where genetic contents differ among individuals in a population. A genome-wide association study (GWAS) seeks to identify genetic variants that contribute to the development and progression of a specific disease.

Association studies discover these genetic factors by correlating an individual's genetic variation with a disease status or disease-related trait. At the genome-wide scale, association studies typically focus on statistical relationships between single-nucleotide polymorphisms (SNPs) and disease traits. SNPs are the most common genetic variants underlying an individual's susceptibility to disease, and associated SNPs are considered to mark the region of a human genome that influences disease risk. A GWAS identifies an SNP as a significant, and therefore associated, variant when the specific genome sequence at the SNP is correlated with a disease trait or disease status. For example, a GWAS may find that individuals with a specific sequence (or allele) at an SNP have higher blood pressure on average than individuals with a different sequence at the SNP. If an SNP has a significant correlation with a trait or disease status, the association study suggests that presence of the particular variant in the neighborhood of the SNP may increase an individual's risk for disease.

Typical analytical strategies for performing association studies rely on standard regression techniques, which assume the data have an identically and independently distributed (i.i.d.) property. If data are i.i.d., all variables are mutually independent because each random variable shares the same probability distribution with other variables. Association study methodology was originally designed for populations composed of unrelated individuals, and standard approaches assume this property is true [6]. However, the big genomic datasets available today inevitably contain distantly related individuals. This genetic relatedness prevents standard association studies from correctly identifying the causal variants and induces identification of many false positive associations (or spurious associations).

Two types of relatedness may produce high rates of false positive associations: ancestry differences and cryptic relatedness. Individuals who share ancestry are more related than individuals from different ancestries. Ancestry differences refer to different ancestry among individuals in a study. Large ($N \geq 5{,}000$) population cohorts inevitably contain individuals who have common ancestry from different populations. Cryptic relatedness exists when some individuals are closely related, but this shared ancestry is unknown to the investigators.

In this review, we refer to population structure as any form of relatedness in the sample, including ancestry differences or cryptic relatedness. These relatedness differences can cause the statistical methodology applied to association studies to assign strong association signals to variants that are not actually causal for the trait or disease. Standard association study techniques applied to population cohorts that contain population structure produce a high rate of false positive associations. These associations may appear to be significant, but they are driven by the cohort's relatedness rather than variants that truly affect trait or disease risk.

Developing GWAS techniques to effectively test for association while correcting for population structure is a computational and statistical challenge. This challenge is relevant to human association studies as well as genetic studies in any organism, including model organisms such as mice. Mouse studies are widely used to study human disease, yet the population structure problem is even more severe than in human studies because the particular history of laboratory mouse strains induced complex patterns of genetic relatedness. These patterns of relatedness can cause false positives in association studies.

Over the past 10 years, new approaches using mixed models have emerged to mitigate the deleterious effects of population structure and relatedness in association studies [7–10]. These approaches were first developed in animal breeding [11] and utilized in the context of model organism studies such as *Arabidopsis* [12, 13] and mouse [8] and were later applied to human studies. In this review, we explicitly characterize population structure as a confounding factor in order to explore the root cause of false positives in association studies. We trace the development of these methods in mouse studies and describe how these methods were adapted to human studies, particularly when they are applied to correct for population structure in large-scale genomic datasets.

Standard GWAS

Genetic association studies attempt to identify SNPs that are responsible for differences in a trait or phenotype values within an individual. An SNP is a single position in the human genome sequence in which individuals in the population have different genetic contents. These differing forms of the same gene are referred to as alleles. SNPs are the most common form of genetic variation, and almost all common SNPs have two alleles. SNPs are ideal targets for association testing [14]. The high level of SNP prevalence suggests they are often correlated with other forms of variation. A single-SNP test measures the correlation between a trait and the genetic information at a specific location in the genome or an SNP. To conduct a typical

single-SNP test, we first collect genetic information at the SNP in a set of individuals (referred to as genotypes). Next, we measure the association (or correlation) of these genotypes with the trait values (or phenotypes) of the individuals (see Fig 1A). This process is repeated for every SNP in the genome. In Fig 1, it is intuitively clear that the first SNP appears to be associated but the second SNP does not appear to be associated.

In order to evaluate whether the association between an SNP and a phenotype is statistically significant, we can use the collected data to test 2 hypotheses. The null hypothesis assumes a model in which the SNP does not affect the phenotype (see Fig 1B). In this hypothesis, the phenotypes (y) are only affected by the population mean (μ) and the environment (e). Unless data indicate otherwise, we assume that the null hypothesis is true and the SNP does not influence the phenotype (i.e., does not affect the individual's disease risk).

An alternative hypothesis provides a model of the SNP being significantly associated with the phenotype (see Fig 1C). In this case, the phenotypes (y) are affected not only by the population mean (μ) and environment (e), but they are also affected by the genotype (x). In other words, the presence of the SNP suggests that an individual is likely to have the trait or disease risk. Here, the quantitative measurement of strength that the genotype has on the phenotype is referred to as the effect size (β). If the effect size (β) is equal to zero, we consider the 2 models equivalent. The SNP is determined to be significantly associated with the phenotype when the data fit the alternative hypothesis beyond a specific threshold.

We mathematically express the null and the alternative hypotheses in order to perform a single-SNP test. We denote the genotype of kth SNP of the jth individual g_{jk} where the genotype is in the set {0,1,2}, which is the number of copies of the kth variant that the jth individual has on their 2 chromosomes. Here, "0" denotes the genotype that does not contain the variant in either chromosome, whereas "1" or "2" denotes the presence of the variant in 1 or 2 of the chromosomes, respectively. In order to simplify the equations for association studies, we standardize the genotypes by subtracting the population mean and dividing by the variance. The frequency of a variant in the population is denoted as p_k, which is the average genotype frequency in the population. The standardized genotypes can be expressed as

$$X_{jk} \in \left\{ \frac{-p_k}{\sqrt{p_k(1-p_k)}}, \frac{1-p_k}{\sqrt{p_k(1-p_k)}}, \frac{2-p_k}{\sqrt{p_k(1-p_k)}} \right\}.$$

Once we have calculated the standardized genotypes, a typical single-SNP test can be used to identify variants associated with traits. A standard regression technique estimates the relationship among variables, including a dependent variable (y), any independent variables (x), and unknown variables (β). Using regression, these simple linear models can correlate the genetic variation with the trait, allowing us to assess whether the data best fit the null or alternative hypothesis.

The equation

$$y_j = \mu + \beta_k X_{jk} + e_j$$

models the phenotype for a single individual j in the study. Here, the effect of the variant on the phenotype is β_k, the model mean is μ, and the contribution of the environment on the phenotype is e_j. The environment's effect on a phenotype for an individual j (e_j) is assumed to be normally distributed with variance σ_e^2, denoted as $e_j \sim N(0, \sigma_e^2)$.

The equation above describes the relationship between the genotype and phenotype of just one individual. We can use vector notation to represent all of the individuals in the dataset and

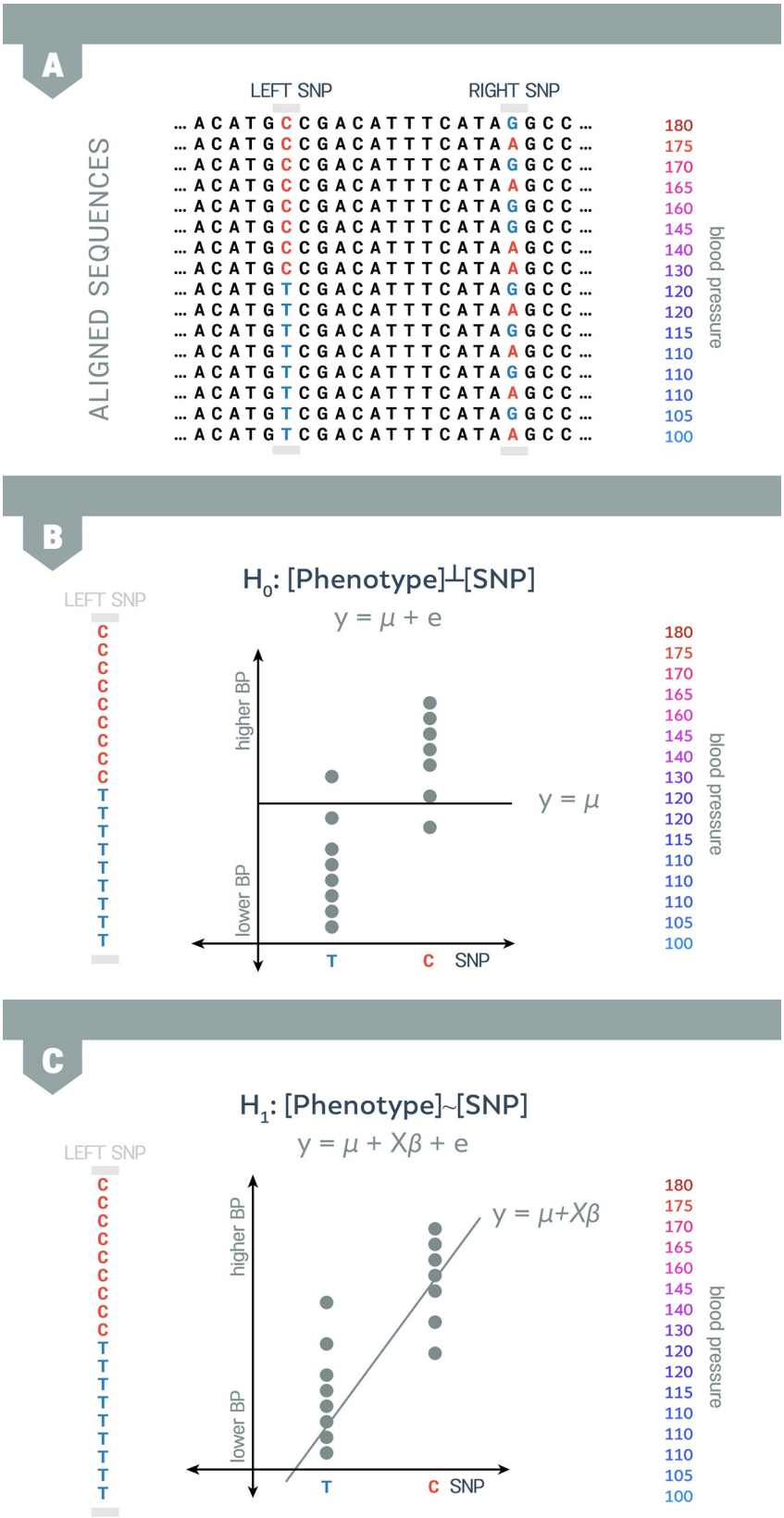
A
LEFT SNP
RIGHT SNP
ALIGNED SEQUENCES
... A C A T G C C G A C A T T T C A T A G G C C ... 180
... A C A T G C C G A C A T T T C A T A A G C C ... 175
... A C A T G C C G A C A T T T C A T A G G C C ... 170
... A C A T G C C G A C A T T T C A T A A G C C ... 165
... A C A T G C C G A C A T T T C A T A G G C C ... 160
... A C A T G C C G A C A T T T C A T A G G C C ... 145
... A C A T G C C G A C A T T T C A T A A G C C ... 140
... A C A T G C C G A C A T T T C A T A A G C C ... 130
... A C A T G T C G A C A T T T C A T A G G C C ... 120
... A C A T G T C G A C A T T T C A T A A G C C ... 120
... A C A T G T C G A C A T T T C A T A G G C C ... 115
... A C A T G T C G A C A T T T C A T A A G C C ... 110
... A C A T G T C G A C A T T T C A T A G G C C ... 110
... A C A T G T C G A C A T T T C A T A A G C C ... 110
... A C A T G T C G A C A T T T C A T A G G C C ... 105
... A C A T G T C G A C A T T T C A T A A G C C ... 100
blood pressure
B
H0: [Phenotype]⊥[SNP]
y = μ + e
LEFT SNP
C C C C C C C C C T T T T T T T T T
higher BP
lower BP
y = μ
T
C
SNP
180 175 170 165 160 145 140 130 120 120 115 110 110 110 105 100
blood pressure
C
H1: [Phenotype]~[SNP]
y = μ + Xβ + e
LEFT SNP
C C C C C C C C C T T T T T T T T T
higher BP
lower BP
y = μ+Xβ
T
C
SNP
180 175 170 165 160 145 140 130 120 120 115 110 110 110 105 100
blood pressure

Fig 1. Standard genetic association study applied to human blood pressure data. (A) The left SNP appears to be more strongly associated with blood pressure than the right SNP. (B) We test 2 hypotheses against each other to evaluate whether the association between an SNP and a phenotype is statistically significant. By default, a null hypothesis assumes that the SNP does not affect the phenotype. (C) If the data fit the alternative hypothesis beyond a certain threshold, the SNP is described as significantly associated with the phenotype. For simplicity, in the diagram, we are depicting only 1 chromosome per individual. BP, blood pressure; SNP, single nucleotide polymorphism.

https://doi.org/10.1371/journal.pgen.1007309.g001

produce the model

$$y = \mu 1 + \beta_k X_k + e, \tag{1}$$

with the phenotypes of all of the individuals in the dataset denoted as a column vector y, a column containing the genotypes for the kth variant in the population denoted as X_k, and a vector containing the environments denoted as e; 1 is a column vector of 1s. We draw the random vector e from the distribution $e \sim N\left(0, \sigma_e^2 I\right)$. We note that each element of e is independent of the others; therefore, the variance–covariance matrix is a diagonal matrix ($\sigma_e^2 I$).

We can write the distribution of y using

$$y \sim N(\mu + \beta_k X_k, \sigma_e^2 I),$$

where I is the identity matrix.

Using the observed data (such as the example in Fig 1), we can estimate the values of the population mean and the effect of the true variant by using the following equations:

$$\hat{\mu} = \frac{\Sigma_{j=1}^{N} y_j}{N} = \frac{1^T y}{N},$$

$$\hat{\beta} = (X_k^T X_k)^{-1} X_k^T y = \frac{\sum_{j=1}^{N} X_{jk} y_j}{N} = \frac{X_k^T y}{N},$$

$$\hat{e} = y - \hat{\mu} 1 - \hat{\beta} X_k,$$

$$\hat{\sigma} = \sqrt{\frac{\sum_{j=1}^{N} \hat{e}_j^2}{N-2}} = \sqrt{\frac{\hat{e}^T \hat{e}}{N-2}},$$

where N is the number of individuals. These equations are simple because the genotypes are standardized. The resulting value is the association between an SNP and a phenotype. We can then test the significance of this association by using the following statistic:

$$S_k = \frac{\hat{\beta}_k}{\hat{\sigma}} \sqrt{N}. \tag{2}$$

This statistic is normally distributed with a mean that depends on the effect of the SNP on the trait, the environmental variance, and the number of individuals. The variance of the statistic is 1. We can write the distribution of the statistic as

$$S_k \sim N\left(\frac{\beta_k}{\sigma_e} \sqrt{N}, 1\right).$$

If the SNP does not have an effect on the trait, the statistic will follow the null distribution,

$$S_k \sim N(0, 1),$$

which is a standard normal distribution. We can then use this null distribution to determine whether the association is significant. This statistic is considered significant with a significance level of α_s if

$$\Phi(S_k) < \alpha_s/2 \text{ or } \Phi(S_k) > 1 - \alpha_s/2,$$

in which case the variant is considered to be associated (see Fig 2). We use the notation α_s to denote the significance level that we need to achieve at any SNP, which in human studies is typically 5×10^{-8} due to multiple testing correction.

The *p*-value of the association is the tail probability area beyond the observed statistic, and the *p*-value can be computed using $2\Phi(-|S_k|)$. If the SNP does not affect the trait, the statistic will come from the null distribution. In this case, the *p*-values will be uniformly distributed between 0 and 1.

An assumed "true" genetic model

Assuming i.i.d., the single-SNP test will tell us whether an SNP is responsible for the differences we observe in an individual's trait or phenotype expression values. However, this simple linear model is an unrealistic model for identifying variants associated with traits in today's large genomic datasets that contain a high degree of relatedness. In real populations, the true effect of a single SNP is influenced by multiple variants that affect the trait. A "hypothetical" true genetic model takes into account the effect of all SNPs on the trait.

Here, the vector notation,

$$y = \mu 1 + \sum_{i=1}^{M} \beta_i X_i + e,$$

models the phenotypes of all the individuals in the dataset, denoted as a column vector *y*. Again, the effect of the *i*th variant on the phenotype is β_i, the mean is μ, and the contribution of the environment on the phenotype is denoted by *e*. Here, the number of variants is *M*.

The true genetic model takes into account the true effect of all SNPs, including the effect of the SNP being tested for association with a trait. When testing SNP *k*, we are using Eq 1, yet

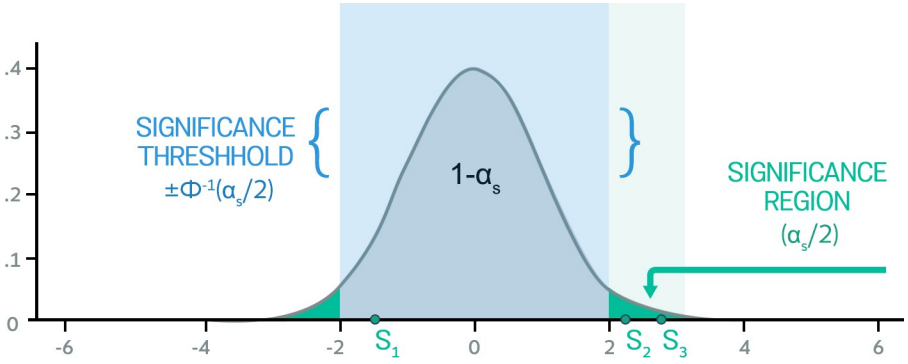

Fig 2. Significance testing in association studies. The null distribution is the standard normal distribution and the expected distribution of the association statistics under the assumption that the effect size is 0. Each variant's association statistic in Eq 2 is computed, and its significance is evaluated using the null distribution. If the statistic falls in the significance region of the distribution, the variant is declared associated. In this example, S1 is not significant, whereas S2 and S3 are significant. The exact location of the threshold is defined as the location on the x-axis where the tail probability area equals the significance threshold (S). This is denoted using the quantile of the standard normal $\Phi^{-1}(x)$.

https://doi.org/10.1371/journal.pgen.1007309.g002

the actual data are generated from

$$y = \mu + X_k\beta_k + \Sigma_{i\neq k}\beta_i X_i + e. \quad (3)$$

In applying the simple linear model to data, we observe a mismatch between the model used for testing and the assumed underlying generative model. Here, any term that is missing in the testing model when compared to the generative model is called an unmodeled factor. The unmodeled factor is exactly $\sum_{i\neq k}\beta_i X_i$.

In this case, the unmodeled factor is the effect of variants in a genome other than the variant being tested. This factor can significantly affect the results of an association study. If the individuals in the study are related to each other, the unmodeled factor may produce a high rate of false positive associations. This is due to the fact that the equations used to estimate the parameters for the model in Eq 1 assume that the covariance matrix of the random vector e is a diagonal matrix, which is not the case in Eq 3.

In an association study, relatedness among individuals is referred to as population structure. Over the past few years, many methods have been developed to mitigate the effects of population structure in association studies. One of the most commonly utilized approaches today, mixed models, originally became popularized in mouse studies and is now the standard approach for analyzing human GWAS studies. In this review, we motivate the problem of population structure in association studies by using laboratory mouse strains and explain how population structure can cause false positive associations. We then motivate mixed models in the context of unmodeled factors.

An example of population structure confounding from mouse genetics

The importance of controlling for population structure is evident in genetic mapping of inbred mouse strains. Mice strains pose particular problems that mixed models are developed to solve, and the basic ideas behind mixed models can be clearly demonstrated with mice genetics. Today's classical inbred laboratory mouse strains descend from a relatively small number of genetic founders (mostly fancy mice originally kept as pets) and are characterized by several population bottlenecks [15, 16].

A second group of laboratory strains are referred to as "wild-derived" strains. These strains include descendants of mice captured in the wild and inbred mice that were never kept as pets. Wild-derived strains do not share the population history of classical laboratory strains. A simple way to visualize the relationship between multiple ancestral groups and traits in the mouse genome is with a phylogenetic tree that can be computed from the genetic information (Fig 3). This tree visualizes the genetic relationships between 32 classical inbred strains and 6 wild-derived strains, using genetic variant information at 140,000 SNPs for each strain.

We observe that the 2 groups of mouse strains are close to each other in the phylogeny and are separated by a long branch length (denoted with a dotted line). This branch represents the many genetic differences between the groups. We also have measurements for body weight and liver weight for each of the 2 strains. Not surprisingly, the body weights of the classical strains are much larger than the body weights of the wild-derived strains (Fig 3). Different selective pressures on the 2 groups, including environmental fitness (wild-derived) and human selection (laboratory), produced these differences in population genetics.

In order to identify which genetic variants are associated with body weight, we applied the linear model described above to 140,000 SNPs from this dataset. In general, we expect association study results to indicate very few significant associations between particular SNPs and a trait. One common way to visualize the results of an association study is with a Manhattan plot. In a Manhattan plot, the mouse genome is plotted against the x-axis, and the measure of significance of correlation between the genome and trait is plotted against the y-axis. Each red

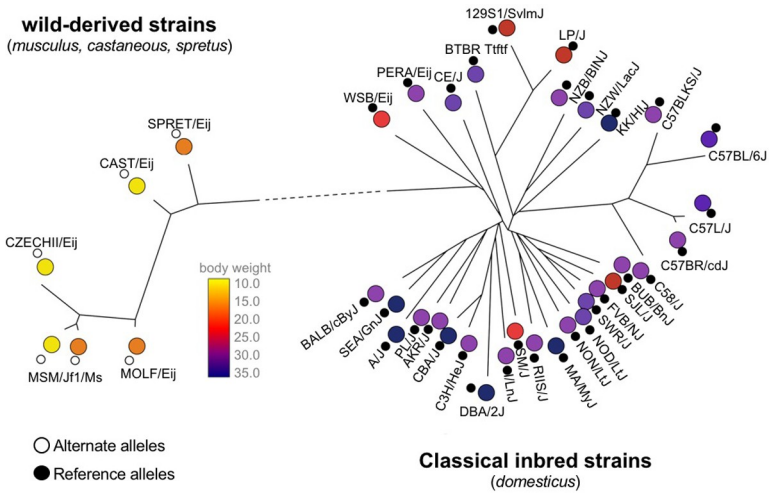

Fig 3. A phylogenetic tree demonstrating the relationships between 38 inbred mouse strains using 140,000 mouse HapMap SNPs. As shown in the tree, the strains cluster in 2 groups: classical inbred strains and wild-derived strains. The body weight phenotypes, obtained from the Mouse Phenome Database, of the strains are shown. Here, classical inbred strains have much higher body weight than wild-derived strains. Many SNPs separate the 2 groups because of the long branch length. One such SNP is shown in the figure. Clearly the SNP is highly correlated with body weight. All of the SNPs that separate these 2 groups will have the same correlation. When we consider both the tree and the SNP, we can infer that the population structure may be driving this correlation and not an effect of the SNP on body weight. SNP, single nucleotide polymorphism.

https://doi.org/10.1371/journal.pgen.1007309.g003

spike represents an SNP at a particular genomic position, and the height of the spike represents the strength of the association with the phenotype. The green horizontal line represents the significance threshold. Any SNP that crosses this line is considered a significant association.

We expect to observe a Manhattan plot similar to the one in Fig 4, in which a number of SNPs affect the phenotypes. Therefore, we would observe signals that cross the threshold at a few locations in the genome, but most of the SNPs will not be associated with the phenotype.

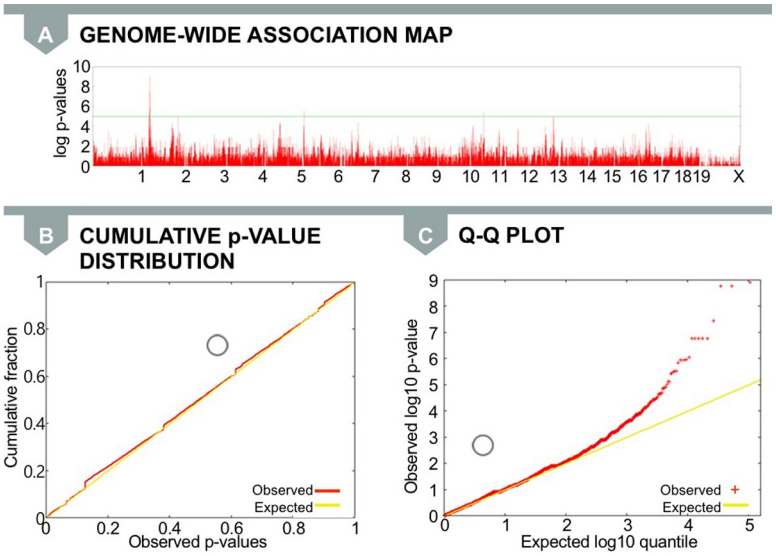

Fig 4. Expected distribution of *p*-values in a typical (A) Manhattan plot, (B) cumulative *p*-value distribution, and (C) Q–Q plot. Circles in (B) and (C) denote where the median *p*-value (red line) falls on the graph in comparison to the expected median *p*-value (yellow line). Here, the median falls close to 0.5, suggesting that population structure is not affecting association results or has been corrected for in the model. Q–Q, quantile–quantile.

https://doi.org/10.1371/journal.pgen.1007309.g004

Another way to visualize the results of an association study is with a cumulative *p*-value distribution plot (Fig 4B) and a quantile-quantile (Q-Q) plot (Fig 4C). These plots are graphical techniques for determining whether multiple datasets come from populations with common distribution. Here, the cumulative *p*-value distribution plot shows the quantiles of the *p*-values, which assess the probable significance of association between a genotype and a trait; the Q–Q plot shows the distribution of the same data log-transformed.

Since we expect most SNPs not be to associated, most of the statistics will be coming from the null distribution. Therefore, most of the *p*-values will be uniformly distributed between 0 and 1. Typically, only a small fraction of the SNPs have signals stronger than expected at the tail of the distribution. This results in a cumulative *p*-value distribution that is close to the diagonal line (Fig 4B) and a Q–Q plot that follows the line for the beginning of the curve (Fig 4C). As shown in Fig 4, we would expect that the median *p*-value would be close to 0.5.

However, when we applied standard linear models to the inbred mouse dataset, we observed strong signals in many locations in the genome (Fig 5A). The cumulative *p*-value distribution and the Q–Q plots are shown in Fig 5B and 5C. In our results, we observe that nearly 50% of the SNPs are significantly associated with the phenotype. There are far more significant associations (red line) than expected associations (yellow line).

Why we observe false positives in mouse genetic studies

We can explain why we observe the excess amount of strong association by examining Fig 3, containing the data for one of the red peaks from the Manhattan plot (found in Fig 5A). Here, the big circles are body weight values, and the small circles represent the variant for each strain for a particular SNP; the black small circle represents the reference allele, and the white small circle represents the alternate allele. When we look at the distribution of body weight values and SNP alleles, it appears that a white allele corresponds to mice with small body weight, whereas a black allele corresponds to mice with large body weight. In Fig 5, there is a very strong correlation between the SNP and the trait of body weight; it is no surprise that we observe a very significant *p*-value.

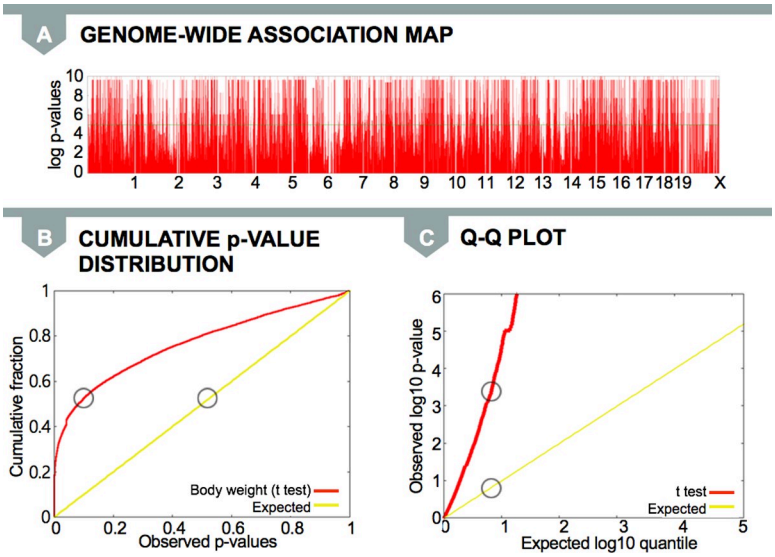

Fig 5. Observed distribution in a (A) Manhattan plot, (B) cumulative *p*-value distribution, and (C) Q–Q plot. Circles in (B) and (C) indicate where the median *p*-value falls on the plot compared to where it is expected. Here, there is a substantial deviation between the red and yellow lines due to inflation of false positive associations for the body weight phenotype. Q–Q, quantile–quantile.

https://doi.org/10.1371/journal.pgen.1007309.g005

However, if we consider the phylogenetic tree over the pattern of alleles at the SNP and body weight values (Fig 3), we see that the separation of the population into classical and wild-derived strains is strongly correlated with the body weight. Here, the SNP differentiates these 2 groups. The length of each tree branch corresponds to the amount of genetic differences between the 2 groups separated by the branch. The long branch length between the classical and wild strains indicates that many SNPs will separate these 2 groups and each one has a strong signal. This correlation between strains and the alleles at these SNPs causes the large amount of observed associations.

Clearly, there are genetic differences between these 2 groups that affect body weight, but not every genetic difference between the 2 groups affects body weight. However, the simple linear model will associate every SNP that separates these 2 groups with body weight. Therefore, most of the associations that we observe are for SNPs that are not actually affecting body weight. These associations are referred to as spurious associations.

Another way to understand the effect of population structure on association is through graphical models. We consider SNPs and traits in Fig 6A. Typically, we perform an association test on an SNP. Observation of an association gives evidence that the SNP affects the trait. On

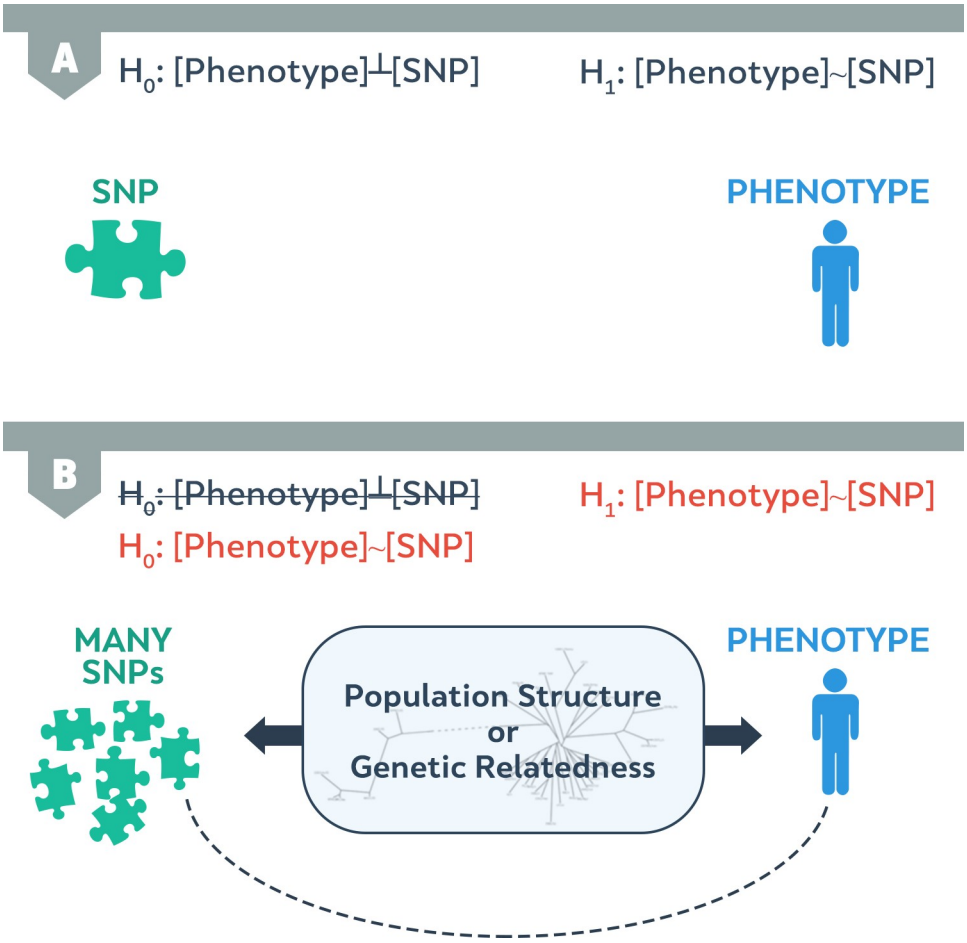

Fig 6. (A) The SNP and the phenotype are independent under the null hypothesis (H_0) and correlated under the alternative hypothesis (H_1). (B) In the case of population structure, the structure will influence many SNPs and the phenotype. In this case, correlation between SNPs and the phenotype will be induced in both the null and alternate hypothesis. SNP, single nucleotide polymorphism.

https://doi.org/10.1371/journal.pgen.1007309.g006

the other hand, if we do not observe an association, this suggests that either the SNP does not affect the trait or that the effect is too small for our study to detect. However, if genetic differences between groups are present (Fig 6B), shared histories will produce many SNPs directly correlated with population structure (straight dark line). In addition, phenotypes, such as body weight, are also highly correlated with the population structure (straight dark line). Such phenotypes will induce correlation between many SNPs and the phenotype (dotted line), including—but not limited to—the SNPs that are actually responsible for variants.

This connection between the phenomenon of false positive associations due to relatedness and unmodeled factors is shown in Eq 3. Here, the genetic history shared between mouse strains is the unmodeled factor $\sum_{i \neq k} \beta_i X_i$. Because the shared genetic history is missing from the testing model, we consider population structure the unmodeled factor.

Using mixed model methods in mouse association studies

We have shown that population structure can bias association study results. Our mouse examples show that we must correct for population structure in order to accurately identify specific genetic variants involved in disease risk. At present, several challenges limit usefulness of genome association studies for implicating genetic variants. First, unmodeled factors are not known and cannot be accounted for in computational methods that match traits with phenotypes. Second, we do not know the exact ways that unmodeled factors interact with population structure to bias output. Finally, many studies ignore dependency among these unmodeled factors.

The effects of these SNPs are the unmodeled factor in Eq 3, and they confound our ability to perform association studies. In reality, there are many SNPs located on the long branching line (Fig 6, dashed line) that affect the phenotype. In order to identify these true associations, we must eliminate the unmodeled factor. Although we cannot know which specific SNPs comprise the unmodeled factor, we can use available knowledge about similarities between the genomes of individuals in our studies to estimate the unmodeled factor.

Using our mouse example, we consider 2 different strains, B6 and C3H. These 2 strains are both classical inbred mice derived from domesticated mice and have similar genomes. In Fig 7A, we show a toy example considering the genomes of the 2 strains. Here, the genomes are very similar; 9 out of 10 SNPs are shared between strains B6 and C3H. In this example, let us assume that the even-numbered SNPs are causal variants that affect the phenotype. For those variants, their corresponding effect size (β_i) will be nonzero. We neither know the actual effect sizes nor the resulting value for the unmodeled factor. However, because they share the same allele at these SNPs, we do know that the 2 strains will have a similar value for the unmodeled factor.

Next, we consider a larger number of strains pairwise (Fig 7B). When we consider the classic inbred mouse strain B6 and the wild mouse strain CAST, the strains have different alleles present at many SNPs. If any of these SNPs affects the trait, the value of the unmodeled factor will differ by the effect size. Therefore, we expect the 2 strains to have different values for the unmodeled factor.

The amount of pairwise sharing of alleles between strains can be used to capture the similarity between the values of the unmodeled factor among strains. In order to do this, we make a matrix that contains all SNPs shared between the paired genomes (Fig 8). This kinship matrix allows us to "model" the values of the unmodeled factors among the individuals in our study, and it shows us which pairs have similar sharing of alleles and which pairs have dissimilar values.

The principle underlying mixed models is that we incorporate this "model" of unmodeled factors into the association test. We incorporate the unknown factors into the model of association using what is called a random effect or a variance component. Our model is called a mixed model because it combines a random effect with the effect sizes of the SNPs we are testing (referred to as fixed effects) to model population structure.

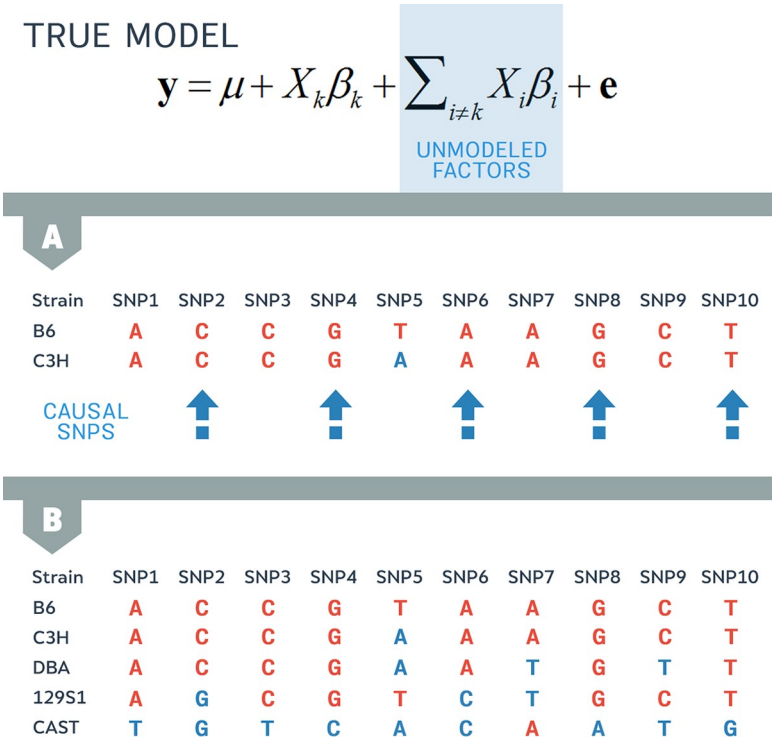

Fig 7. Pairwise similarity between strains gives some insight into the similarity of the unmodeled factor. In this toy example, we consider 10 SNPs in which the even-numbered SNPs are the causal SNPs with an effect on the trait. (A) Because B6 and C3H share alleles at 9 out of 10 SNPs, these strains have a similar value for the unmodeled factor. (B) When we consider other strains, the unmodeled factors may be larger. For example, B6 and CAST, which share few SNPs, will have different values for their unmodeled factor. SNP, single nucleotide polymorphism.

https://doi.org/10.1371/journal.pgen.1007309.g007

When using a mixed model to identify causal variation, one key step is to establish these fixed parameters and random effect components. A linear mixed model (LMM) uses the information from the matrix to account for the unmodeled factor. We extend the simple, hypothetical true model,

$$y = \mu 1 + \beta_k X_k + e,$$

to include a term that captures the unmodeled factors. The term u in

$$y = \mu 1 + \beta_k X_k + u + e \quad (4)$$

is a random vector that depends on the amount of shared genome in terms of pairwise differences. Here, we assume that $u \sim N\left(0, \sigma_g^2 K\right)$, where K is the kinship matrix. Each entry of K estimates the pairwise similarity between the genomes of the individuals in the study, which follows the intuition of Figs 7 and 8.

In practice, K can be computed from the genotypes in which each entry in the kinship matrix is just the product of the standardized genotypes for the 2 individuals divided by the number of variants. Therefore, the kinship entry computing the relatedness between

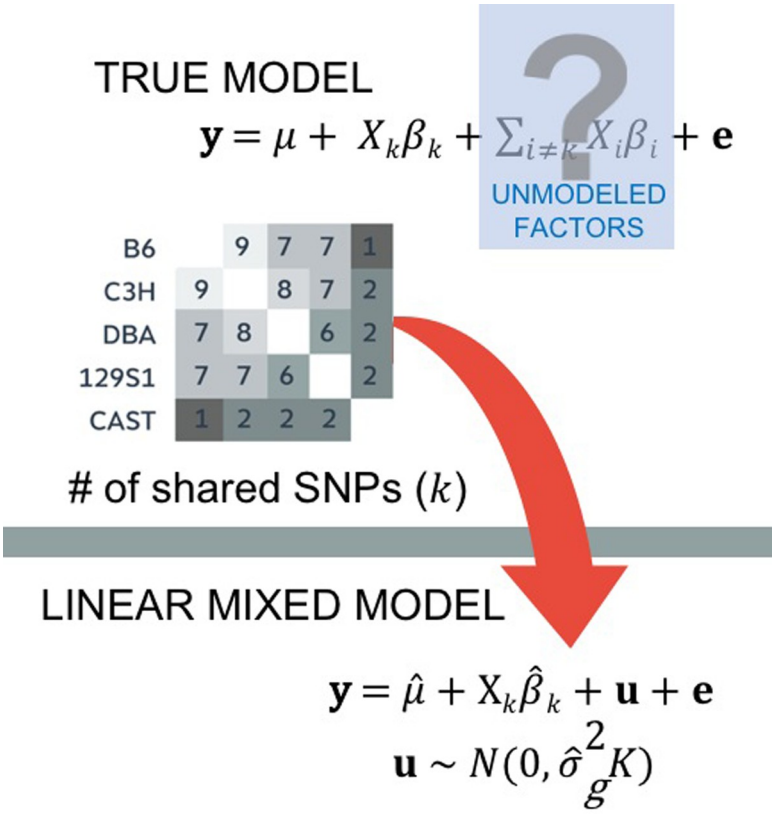

Fig 8. The mixed model includes a term u which attempts to model the unmodeled factors in the true model. The term uses information from the kinship matrix that accounts for the dependency among SNPs correlated with phenotypes due to population structure. SNP, single nucleotide polymorphism.

https://doi.org/10.1371/journal.pgen.1007309.g008

individuals i and j is

$$K_{ij} = \frac{\sum_{k=1}^{M} X_{ik} X_{jk}}{M}.$$

We can elegantly compute the kinship matrix using the equation $K = XX^T/M$.

The mixed model is making an assumption that the phenotype follows the model in Eq 4. How accurately this assumption holds in practice depends on how well the kinship matrix captures the covariance between individuals for the phenotype. Exploring this issue is an active area of research, leading to many variations of mixed models including techniques for computing kinship matrices. We discuss some of these developments later in this review.

The standard estimation equations above cannot be used to estimate the values of the parameters in Eq 4. Due to the random effect u, the phenotypes of the individuals are no longer independent of each other—an assumption of the previous methods.

However, if we know the values of σ_g^2 and σ_e^2, we can apply the following "mixed model trick." We note that the phenotypes will follow the distribution,

$$y \sim N(\mu + \sum \beta_i X_i, V),$$

where $V = \sigma_g^2 K + \sigma_e^2 I$ and I is the identity matrix. If we transform then multiply the

phenotypes and genotypes by $V^{-\frac{1}{2}}$, we get

$$V^{-\frac{1}{2}}y \sim N\left(V^{-\frac{1}{2}}1\mu + \sum \beta V^{-\frac{1}{2}}X_i, I\right).$$

In the transformed data, the individuals are now independent of each other, and we can apply the estimation equations presented above to estimate the values for β and the association statistics.

In this case, we assume that the β_i values are drawn from a normal distribution with a mean zero as effect size and σ_e^2 as the variance.

Estimating the values of σ_g^2 and σ_e^2 is a difficult computational problem referred to as estimating the variance components. These parameters are estimated by utilizing a maximum likelihood approach. Specifically, we attempt to find the values of σ_g^2 and σ_e^2, such that the following log likelihood function of the data is maximized:

$$l(y, X_k, \beta_k, \sigma_g, \sigma_e, n) = -\frac{1}{2}[n\log(2\pi) + \log \mid V \mid + (y - X_k\beta_k)V^{-1}(y - X_k\beta_k)],$$

where $V = \sigma_g^2 K + \sigma_e^2 I$.

This equation is computationally difficult because likelihood requires computing the inverse of the matrix (V^{-1}), which in turn depends on the values of σ_g and σ_e. Optimization methods that maximize this likelihood apply algorithms updating current estimates of σ_g and σ_e until they converge to high values of the log likelihood function. Each step of an optimization algorithm is referred to as an iteration. In each iteration, the optimization algorithm must evaluate the log likelihood for the current values of σ_g and σ_e and must compute this matrix inverse. A straightforward way to compute a matrix inverse involves a complexity of approximately $O(N^3)$. Unfortunately, this results in a very inefficient algorithm and prevents mixed models from being widely utilized in association studies, despite their long history in genetics.

The idea to infer relationships from the SNPs directly was originally proposed in [8], yet mixed models were used in genetics for decades prior to where the kinship matrix was inferred from pedigrees. In the same paper, we also presented efficient mixed model association (EMMA) [8], an efficient algorithm for estimating these parameters. Since we first presented EMMA, many other groups have developed similar efficient algorithms [7, 9, 17]. The key idea behind EMMA is that we apply spectral decomposition to the kinship matrix, leading to a much faster optimization algorithm. This type of approach was previously utilized in classical mixed model approaches applied to pedigrees. The spectral decomposition only needs to be computed once and requires a complexity of $O(N^3)$. Specifically, if we write $K = UDU^T$, where U is a matrix of eigenvectors and D is a diagonal matrix of eigenvalues, then we can represent V using matrix algebra properties as follows:

$$V = \sigma_g^2 K + \sigma_e^2 I = \sigma_g^2 UDU^T + \sigma_e^2 UIU^T = U(\sigma_g^2 D + \sigma_e^2 I)U^T.$$

We can then compute the quantity $z = U^T(y - X_k\beta_k)$ for each SNP k, which has complexity $O(N^2)$. The log likelihood of the data can then be computed using

$$l(y, X_k, \beta_k, \sigma_g, \sigma_e, n) = -\frac{1}{2}[n\log(2\pi) + \sigma_g^2 \mathrm{Tr}(D) + n\sigma_e^2 + z^T(\sigma_g^2 D + \sigma_e^2 I)^{-1} z],$$

which can be computed in complexity $O(N)$ because the matrix inside the likelihood is now diagonal. The inverse can be computed by simply taking the reciprocal of the elements along the diagonal. This procedure results in a very efficient algorithm that is useful for today's large-scale human genomic datasets.

We applied EMMA to the same mouse association data that we analyzed using a standard LMM approach (see Fig 3). With these computational improvements, we almost completely reduced the inflation of false positives while obtaining nearly uniform *p*-value distribution for most SNPs (Fig 9). Here, the strongest peak, which is not significant, falls into a region of the genome on chromosome 8, which is known to be associated with body weight. Regions of the genome that correlate with variation in a phenotype are referred to as quantitative trait loci (QTL).

Next, we applied EMMA to other phenotypes from the same mouse strain datasets, including a liver weight phenotype. Here, we see that the inflation of false positives is reduced and a strong signal at chromosome 2 (chr2) is more pronounced after the correction (Fig 10). EMMA correctly identifies a locus for liver weight that falls into the QTL Lvrq1 (liver weight), which was previously identified using a traditional mouse mapping approach [18].

Population structure and mixed models in human association studies

When mixed models were first used in mouse studies, the problem of relatedness in human GWAS studies was a well-known challenge. At that time, there was no single approach to handle relatedness. Instead, different types of relatedness were explicitly modeled, and association study methods were adapted to those scenarios. There is an entire class of methods designed to handle relatedness when there are closely related individuals in the genetic study and the genetic relationships are known. These include methods for multigenerational families, twins, and siblings [19, 20].

A complication in human association studies is when the relationships are unknown. One of the most common types of relatedness among individuals in human studies is due to ancestry. Ancestry refers to the population that an individual descended from. Many individuals are admixed, which means they are descended from ancestors in different populations. If an association study contains individuals from different populations or differing degrees of admixture, the individuals will have different degrees of relatedness among them. In other words, individuals with the same ancestry are slightly more related to each other than individuals with different ancestries.

It is well documented that these ancestry differences can induce false positive associations [21]. Association studies that analyzed individuals with differences in ancestry typically utilized an approach to predict the ancestry for each individual and then incorporated this

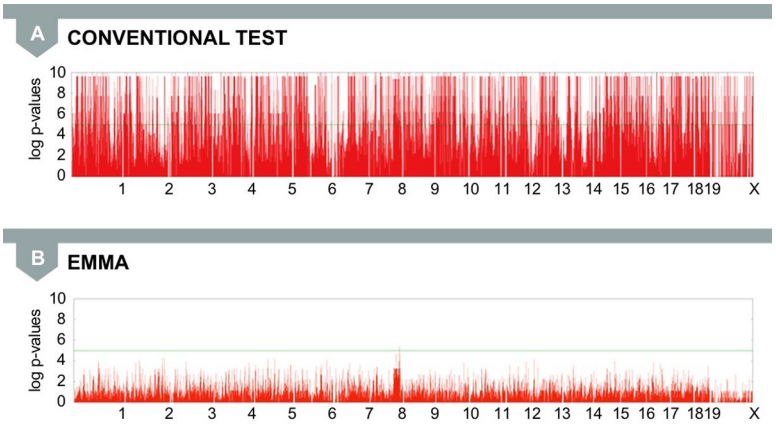

Fig 9. (A) The conventional GWAS test applied to mouse body weight phenotypes produces numerous false positives. (B) The mixed model approach using EMMA almost completely reduces the inflation of false positives and identifies a strong peak (chr8) that falls into a known body weight QTL. chr8, chromosome 8; EMMA, efficient mixed model association; GWAS, genome-wide association study; QTL, quantitative trait loci.

https://doi.org/10.1371/journal.pgen.1007309.g009

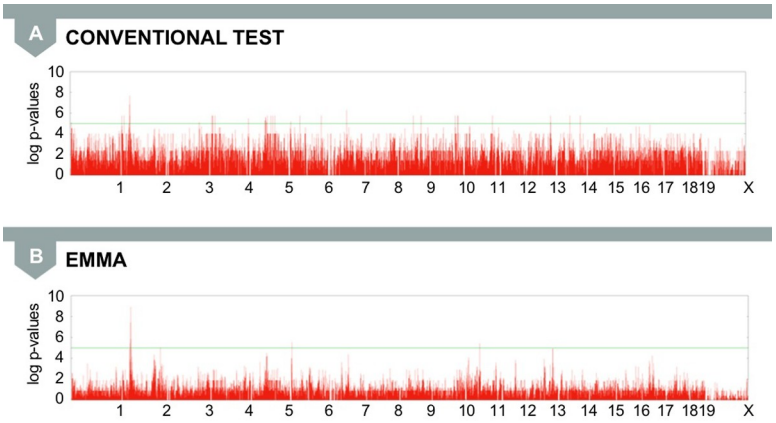

Fig 10. (A) The conventional GWAS test applied to mouse liver weight phenotypes produces numerous false positive associations. (B) The mixed model approach using EMMA reduces inflation of false positives and correctly produces a stronger signal at chr2, a region that is located in known QTLs for liver weight. chr2, chromosome 2; EMMA, efficient mixed model association; GWAS, genome-wide association study; QTL, quantitative trait loci.

https://doi.org/10.1371/journal.pgen.1007309.g010

information as a covariate in the model [22]. An alternate approach was to estimate principal components over the genotype data, which could be interpreted as a proxy for ancestry information and included in the model as covariates [23]. In the human genetics literature, ancestry differences are sometimes referred to as population structure. In this review, we use the term ancestry differences separately from the term population structure; we use the latter to specifically describe the general phenomenon of relatedness in a sample.

A second type of relatedness is cryptic relatedness [24]. Because GWASs are applied to extremely large samples, there are often individuals included in the study who happen to be related, but this relatedness is unknown to both the individuals and the investigators. Typically, cryptic relatedness is handled by screening the association study for related individuals and computing the genetic similarity between each pair of individuals.

A general purpose approach to correct for population structure, or any type of confounding in association studies, is genomic control [25, 26]. Genomic control allows us to measure the extent to which population structure (or any other confounder) is affecting the association statistics. By examining the cumulative p-value distribution plot, we consider the deviation of the actual plot from what is expected at the median. Because we expect the vast majority of variants not to be associated with the trait, we expect the median observed p-value to be close to 0.5. Typically, population structure induces a more significant observed median p-value.

Genomic control computes a correction factor referred to as λ, which is a scaling factor used to scale all observed p-values so the corrected median p-value will be 0.5. The λ is on the χ^2 scale (meaning that the median p-value is converted to a χ^2 value and the ratio is computed relative to the χ^2 value) corresponding to a p-value of 0.5, which is 0.545. The observed association p-values are converted from p-values to χ^2 statistics, scaled by λ, and then converted back to p-values.

We can also use the value of the λ as a measure of the extent of the effect of confounding on the association statistics. Genomic control λ's are widely utilized to compare different correction approaches. A λ of 1.0 shows that there is no inflation. A value greater than 1.0 is evidence that the association statistics are inflated. Typically, the 95% confidence interval of the λ in GWASs is 0.02. Therefore, any λ of 1.03 or higher suggests that there is some inflation. We note that more recent exploration of polygenicity, or the amount of causal variants for a trait, suggests that there are many more causal variants than originally expected. In this case, the λ

values should actually be higher than 1.0 [27]. We discuss this perspective in the Discussion section of this paper.

In the literature, ancestry differences and cryptic relatedness are referred to as distinct phenomenon. In fact, they can be thought of as different degrees of relatedness in the sample. Consider Fig 11A, which shows a potential pedigree relating all of the individuals in an association study sample. Cryptic relatedness can be thought of as relatedness in a more recent portion of the tree (Fig 11B), and ancestry differences can be thought of as relatedness near the top of the tree (Fig 11C).

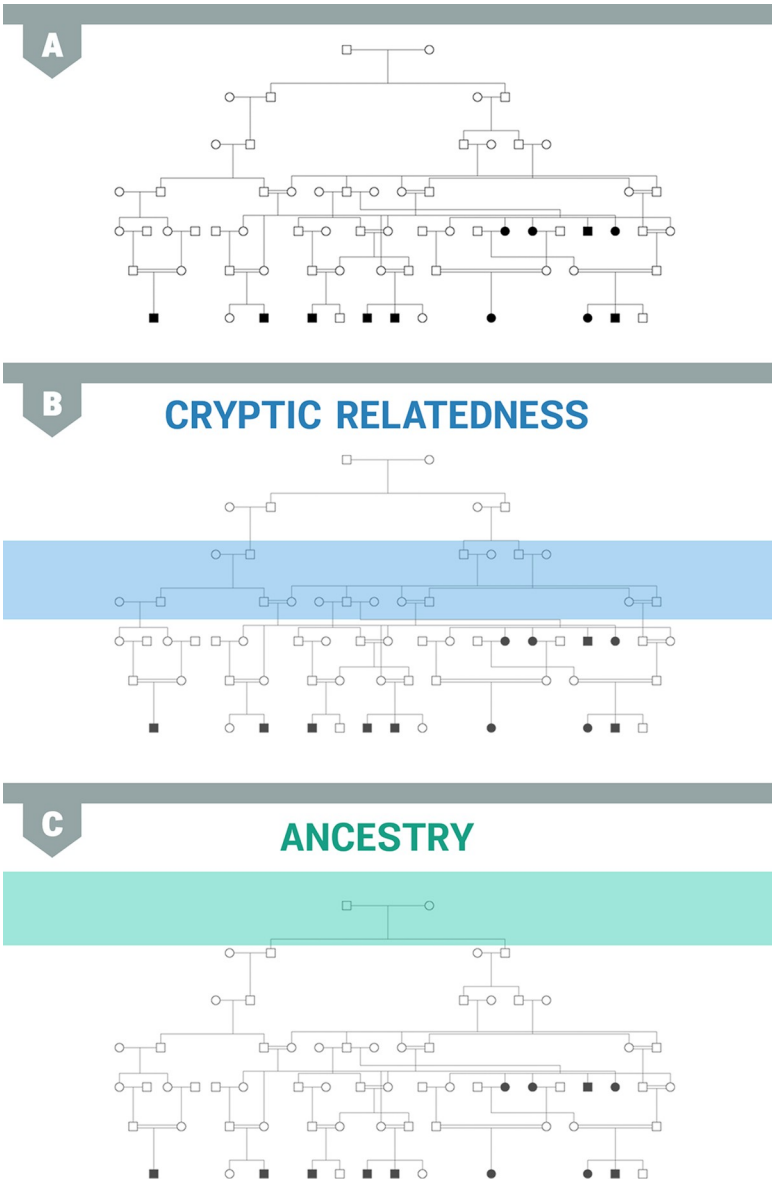

Fig 11. Different degrees of relatedness in the sample. (A) All of the individuals in a genetic study are somehow related through a large pedigree or family tree. Different parts of the tree induce different types of relatedness. (B) Cryptic relatedness refers to relatively recent genetic relationships. (C) Relatedness due to ancestry refers to relatedness caused by ancestors originating from the same region. The boxes in (B) and (C) represent the level of the pedigree that causes that type of relatedness in each case, respectively.

https://doi.org/10.1371/journal.pgen.1007309.g011

Mixed models can handle nearly arbitrary genetic relationships between individuals and are a natural approach for human association studies. Mixed models are ideal because they can be applied without explicit identification of the relatedness within the sample. They also enable the analysis of datasets with particularly complex genetic relationships, such as isolate populations in which the population is descended from a small number of founder individuals [28]. For isolate populations, the previous methods were not able to fully account for population structure.

Mixed models were first used in human studies with the Northern Finnish Birth Cohort [29], using a version of the EMMA software called EMMAX [9], in which mixed models were applied to 331,475 SNPs in 5,326 individuals who were phenotypes for 10 traits. These traits include C-reactive protein (CRP), triglyceride (TG), insulin plasma levels (INS), diastolic blood pressure (DBP), body mass index (BMI), glucose (GLU), high-density lipoprotein (HDL), systolic blood pressure (SBP), and low-density lipoprotein A (LDL). Individuals within this cohort have some ancestry differences due to their origin from different parts of Finland, and they share some genetic relationships.

Table 1 shows the results of applying mixed models to these traits. Each entry in the table shows the λ value for the analysis of that phenotype. The first column shows the results of the uncorrected analysis. We can see that there are very large λ factors, particularly for height. In fact, the associations with height were not reported in the original Sabatti and colleagues [29] manuscript because the high λ value suggested that some of the observed associations may be false positives. The second column shows the λ factors after eliminating cryptically related individuals. Here, we computed the pairwise relationships between individuals and filtered out one of any pair that was closely related. This approach filtered out 611 individuals.

The third column shows the λ factors after using the principal component approach to correct for ancestry differences. One hundred principal components were utilized in this analysis, which is a much larger number of components than are typically utilized to show the limit of the approach. The last column shows the λ for mixed models. Each of these λ values is within the 95% confidence interval (around 1.0), suggesting that mixed models can correct for all of the population structure in the sample—including cryptic relatedness and ancestry differences. As shown in Table 1, only mixed models adequately correct for population structure in this sample.

Discussion and recent developments

Over the past decade, association studies have identified thousands of variants implicated in dozens of common human diseases. The traditional approach to association studies assumes

Table 1. Results of analysis (λ values) on NFBC66 data.

Traits	Uncorrected	IBD < 0.1	100PC	EMMAX
BMI	1.036	1.028	1.024	1.001
CRP	1.012	1.020	1.020	0.994
DBP	1.033	1.025	1.029	1.010
GLU	1.045	1.025	1.030	1.009
HDL	1.054	1.041	1.037	1.003
INS	1.026	1.026	1.015	1.005
LDL	1.093	1.089	1.040	1.002
SBP	1.063	1.054	1.021	1.004
TG	1.024	1.021	1.018	0.999
Height	1.193	1.152	1.080	1.002

Abbreviations: BMI, body mass index; CRP, C-reactive protein; DBP, diastolic blood pressure; GLU, glucose; HDL, high-density lipoprotein; INS, insulin plasma levels; LDL, low-density lipoprotein A; NFBC66, Northern Finish Birth Cohort 66; SBP, systolic blood pressure;TG, triglyceride.

https://doi.org/10.1371/journal.pgen.1007309.t001

that individuals are unrelated to each other. However, in practice, individuals in genetic studies are related to each other in numerous complex ways. In this review, we demonstrate how these relationships cause false positives in association studies and how mixed models can correct for these confounding genetic relationships.

This review covers only the basic principles of mixed models and population structure. Since the original EMMA paper in 2008, mixed models have become an active research area. Many groups have published papers exploring various aspects of mixed models and their application to complex genomic problems.

For example, many approaches have been developed to improve the efficiency of mixed models, including the methods Fast-LMM [17] and GEMMA [7]. More recently, a method called BOLT-LMM [30] was developed for scaling analyses to handle cohorts in the hundreds of thousands of individuals.

Another direction of method development has been extending mixed models to handle case control studies. These approaches typically assume a liability threshold model in which there is an underlying continuous phenotype; if the phenotype is above a threshold, the individual has a disease. If it is below a threshold, the individual does not have the disease [31]. These types of studies are also complicated by the phenomenon of selection bias because the cases are oversampled from the population. At present, such mixed model extensions to case and/or control studies result in challenging computational problems [32–34].

Another direction in mixed models research is based on observations that a bias is induced when the SNP that is tested is also used in the computation of the kinship matrices [10]. This bias motivated the idea that, when applying mixed models, the kinship matrix should not contain the SNP being tested. As a result, the Leave One Chromosome Out (LOCO) approach constructs a different kinship matrix for testing each chromosome and leaves out the SNPs on the chromosome being tested [35]. Methods incorporating mixed models that use LOCO have higher statistical power compared to traditional association studies.

Our example in Fig 8 brings to the surface a key issue related to the application of mixed models in genetic studies. In our example, all of the variants are used to build the kinship matrix, yet only a subset of them are the actual causal variants affecting the trait. The model also makes assumptions about the magnitude of the contribution of each SNP to the trait. Issues related to these assumptions have been explored in depth in the literature [36]. Proposed approaches include stratifying the variants based on frequency when constructing the kinship matrix [37–39] and taking into account linkage disequilibrium when generating the kinship matrix [40, 41]. In general, whether or not mixed model approaches developed for common variants will be effective for rare variants is an active area of current research [42].

The results of GWASs have demonstrated that many complex traits are highly polygenic, suggesting that there are hundreds (if not thousands) of loci that influence some traits [43]. Some traits, such as height, are known to be highly polygenic. In this case, it is not clear what the actual value of λ should be for a polygenic trait as it is expected to have a contribution from both confounding effects as well as polygenicity. More recently, a method called LD (Linkage Disequilibrium) score regression has been developed that attempts to differentiate between these 2 components [44].

Mixed models are also utilized in genetic studies beyond mere correction for population structure as described in this review. Mixed models have become important in human GWAS analysis because the estimates of σ_g^2 and σ_e^2 can be used to estimate the heritability of the trait. Recent results suggest that common variants explain a larger proportion of the variance of complex traits than previously thought [2, 4, 45]. Common variants are also utilized to capture environmental factors that may be correlated with genetic background [46] and even to model

gene-by-environment interactions [47]. Mixed models are also utilized to understand the proportion of regions of the genome or types of functional elements that contribute to a trait [43, 48, 49].

From their origins in analyzing genetic variation in nonhuman organisms to powering large-scale human GWASs today, mixed models play an important role in the analysis of genetic data, particularly in correcting for population structure. Improving and extending mixed model approaches is now an active area of research in human genomics.

References

1. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009; 461(7265):747–53. Epub 2009/10/09. https://doi.org/10.1038/nature08494 PMID: 19812666; PubMed Central PMCID: PMCPMC2831613.
2. International Schizophrenia C, Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009; 460(7256):748–52. Epub 2009/07/03. https://doi.org/10.1038/nature08185 PMID: 19571811; PubMed Central PMCID: PMCPMC3912837.
3. Stram DO. Design, analysis, and interpretation of genome-wide association scans. New York: Springer; 2014. xv, 334 pages p.
4. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010; 42(7):565–9. Epub 2010/06/22. https://doi.org/10.1038/ng.608 PMID: 20562875; PubMed Central PMCID: PMCPMC3232052.
5. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017; 45(D1):D896–D901. Epub 2016/12/03. https://doi.org/10.1093/nar/gkw1133 PMID: 27899670; PubMed Central PMCID: PMCPMC5210590.
6. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science. 1996; 273 (5281):1516–7. Epub 1996/09/13. PMID: 8801636.
7. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012; 44(7):821–4. Epub 2012/06/19. https://doi.org/10.1038/ng.2310 PMID: 22706312; PubMed Central PMCID: PMCPMC3386377.
8. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, et al. Efficient control of population structure in model organism association mapping. Genetics. 2008; 178(3):1709–23. Epub 2008/04/04. https://doi.org/10.1534/genetics.107.080101 PMID: 18385116; PubMed Central PMCID: PMCPMC2278096.
9. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010; 42(4):348–54. Epub 2010/03/09. https://doi.org/10.1038/ng.548 PMID: 20208533; PubMed Central PMCID: PMCPMC3092069.
10. Listgarten J, Lippert C, Kadie CM, Davidson RI, Eskin E, Heckerman D. Improved linear mixed models for genome-wide association studies. Nat Methods. 2012; 9(6):525–6. Epub 2012/06/07. https://doi.org/10.1038/nmeth.2037 PMID: 22669648; PubMed Central PMCID: PMCPMC3597090.
11. Henderson CR. Best linear unbiased estimation and prediction under a selection model. Biometrics. 1975; 31(2):423–47. Epub 1975/06/01. PMID: 1174616.
12. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet. 2006; 38(2):203–8. Epub 2005/12/29. https://doi.org/10.1038/ng1702 PMID: 16380716.
13. Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, et al. An Arabidopsis example of association mapping in structured samples. PLoS Genet. 2007; 3(1):e4. Epub 2007/01/24. https://doi.org/10.1371/journal.pgen.0030004 PMID: 17238287; PubMed Central PMCID: PMCPMC1779303.
14. International HapMap C, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010; 467(7311):52–8. Epub 2010/09/03. https://doi.org/10.1038/nature09298 PMID: 20811451; PubMed Central PMCID: PMCPMC3173859.
15. Frazer KA, Eskin E, Kang HM, Bogue MA, Hinds DA, Beilharz EJ, et al. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. Nature. 2007; 448(7157):1050–3. Epub 2007/07/31. https://doi.org/10.1038/nature06067 PMID: 17660834.

16. Yang H, Bell TA, Churchill GA, Pardo-Manuel de Villena F. On the subspecific origin of the laboratory mouse. Nat Genet. 2007; 39(9):1100–7. Epub 2007/07/31. https://doi.org/10.1038/ng2087 PMID: 17660819.

17. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear mixed models for genome-wide association studies. Nat Methods. 2011; 8(10):833–5. Epub 2011/09/06. https://doi.org/10.1038/nmeth.1681 PMID: 21892150.

18. Rocha JL, Eisen EJ, Van Vleck LD, Pomp D. A large-sample QTL study in mice: I. Growth. Mamm Genome. 2004; 15(2):83–99. Epub 2004/04/03. PMID: 15058380.

19. Freimer N, Sabatti C. The use of pedigree, sib-pair and association studies of common diseases for genetic mapping and epidemiology. Nat Genet. 2004; 36(10):1045–51. Epub 2004/09/30. https://doi.org/10.1038/ng1433 PMID: 15454942.

20. van Dongen J, Slagboom PE, Draisma HH, Martin NG, Boomsma DI. The continuing value of twin studies in the omics era. Nat Rev Genet. 2012; 13(9):640–53. Epub 2012/08/01. https://doi.org/10.1038/nrg3243 PMID: 22847273.

21. Helgason A, Yngvadottir B, Hrafnkelsson B, Gulcher J, Stefansson K. An Icelandic example of the impact of population structure on association studies. Nat Genet. 2005; 37(1):90–5. Epub 2004/12/21. https://doi.org/10.1038/ng1492 PMID: 15608637.

22. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. Am J Hum Genet. 2000; 67(1):170–81. Epub 2000/05/29. https://doi.org/10.1086/302959 PMID: 10827107; PubMed Central PMCID: PMCPMC1287075.

23. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38(8):904–9. Epub 2006/07/25. https://doi.org/10.1038/ng1847 PMID: 16862161.

24. Voight BF, Pritchard JK. Confounding from cryptic relatedness in case-control association studies. PLoS Genet. 2005; 1(3):e32. Epub 2005/09/10. https://doi.org/10.1371/journal.pgen.0010032 PMID: 16151517; PubMed Central PMCID: PMCPMC1200427.

25. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999; 55(4):997–1004. Epub 2001/04/21. PMID: 11315092.

26. Bacanu SA, Devlin B, Roeder K. Association studies for quantitative traits in structured populations. Genet Epidemiol. 2002; 22(1):78–93. Epub 2002/01/05. https://doi.org/10.1002/gepi.1045 PMID: 11754475.

27. Yang J, Weedon MN, Purcell S, Lettre G, Estrada K, Willer CJ, et al. Genomic inflation factors under polygenic inheritance. Eur J Hum Genet. 2011; 19(7):807–12. Epub 2011/03/17. https://doi.org/10.1038/ejhg.2011.39 PMID: 21407268; PubMed Central PMCID: PMCPMC3137506.

28. Kenny EE, Kim M, Gusev A, Lowe JK, Salit J, Smith JG, et al. Increased power of mixed models facilitates association mapping of 10 loci for metabolic traits in an isolated population. Hum Mol Genet. 2011; 20(4):827–39. Epub 2010/12/02. https://doi.org/10.1093/hmg/ddq510 PMID: 21118897; PubMed Central PMCID: PMCPMC3024042.

29. Sabatti C, Service SK, Hartikainen AL, Pouta A, Ripatti S, Brodsky J, et al. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. Nat Genet. 2009; 41(1):35–46. Epub 2008/12/09. https://doi.org/10.1038/ng.271 PMID: 19060910; PubMed Central PMCID: PMCPMC2687077.

30. Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjalmsson BJ, Finucane HK, Salem RM, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet. 2015; 47(3):284–90. Epub 2015/02/03. https://doi.org/10.1038/ng.3190 PMID: 25642633; PubMed Central PMCID: PMCPMC4342297.

31. Zaitlen N, Lindstrom S, Pasaniuc B, Cornelis M, Genovese G, Pollack S, et al. Informed conditioning on clinical covariates increases power in case-control association studies. PLoS Genet. 2012; 8(11):e1003032. Epub 2012/11/13. https://doi.org/10.1371/journal.pgen.1003032 PMID: 23144628; PubMed Central PMCID: PMCPMC3493452.

32. Hayeck TJ, Zaitlen NA, Loh PR, Vilhjalmsson B, Pollack S, Gusev A, et al. Mixed model with correction for case-control ascertainment increases association power. Am J Hum Genet. 2015; 96(5):720–30. Epub 2015/04/22. https://doi.org/10.1016/j.ajhg.2015.03.004 PMID: 25892111; PubMed Central PMCID: PMCPMC4570278.

33. Golan D, Rosset S. Effective genetic-risk prediction using mixed models. Am J Hum Genet. 2014; 95(4):383–93. Epub 2014/10/04. https://doi.org/10.1016/j.ajhg.2014.09.007 PMID: 25279982; PubMed Central PMCID: PMCPMC4185122.

34. Weissbrod O, Lippert C, Geiger D, Heckerman D. Accurate liability estimation improves power in ascertained case-control studies. Nat Methods. 2015; 12(4):332–4. Epub 2015/02/11. https://doi.org/10.1038/nmeth.3285 PMID: 25664543.

35. Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. Nat Genet. 2014; 46(2):100–6. Epub 2014/01/30. https://doi.org/10.1038/ng.2876 PMID: 24473328; PubMed Central PMCID: PMCPMC3989144.

36. Yang J, Zeng J, Goddard ME, Wray NR, Visscher PM. Concepts, estimation and interpretation of SNP-based heritability. Nat Genet. 2017; 49(9):1304–10. Epub 2017/08/31. https://doi.org/10.1038/ng.3941 PMID: 28854176.

37. Purcell SM, Moran JL, Fromer M, Ruderfer D, Solovieff N, Roussos P, et al. A polygenic burden of rare disruptive mutations in schizophrenia. Nature. 2014; 506(7487):185–90. Epub 2014/01/28. https://doi.org/10.1038/nature12975 PMID: 24463508; PubMed Central PMCID: PMCPMC4136494.

38. Marouli E, Graff M, Medina-Gomez C, Lo KS, Wood AR, Kjaer TR, et al. Rare and low-frequency coding variants alter human adult height. Nature. 2017; 542(7640):186–90. Epub 2017/02/02. https://doi.org/10.1038/nature21039 PMID: 28146470; PubMed Central PMCID: PMCPMC5302847.

39. Speed D, Cai N, Consortium U, Johnson MR, Nejentsev S, Balding DJ. Reevaluation of SNP heritability in complex human traits. Nat Genet. 2017; 49(7):986–92. Epub 2017/05/23. https://doi.org/10.1038/ng.3865 PMID: 28530675; PubMed Central PMCID: PMCPMC5493198.

40. Speed D, Hemani G, Johnson MR, Balding DJ. Improved heritability estimation from genome-wide SNPs. Am J Hum Genet. 2012; 91(6):1011–21. Epub 2012/12/12. https://doi.org/10.1016/j.ajhg.2012.10.010 PMID: 23217325; PubMed Central PMCID: PMCPMC3516604.

41. Yang J, Lee SH, Wray NR, Goddard ME, Visscher PM. GCTA-GREML accounts for linkage disequilibrium when estimating genetic variance from genome-wide SNPs. Proc Natl Acad Sci U S A. 2016; 113(32):E4579–80. Epub 2016/07/28. https://doi.org/10.1073/pnas.1602743113 PMID: 27457963; PubMed Central PMCID: PMCPMC4987770.

42. Mathieson I, McVean G. Differential confounding of rare and common variants in spatially structured populations. Nat Genet. 2012; 44(3):243–6. Epub 2012/02/07. https://doi.org/10.1038/ng.1074 PMID: 22306651; PubMed Central PMCID: PMCPMC3303124.

43. Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, et al. Genome partitioning of genetic variation for complex traits using common SNPs. Nat Genet. 2011; 43(6):519–25. Epub 2011/05/10. https://doi.org/10.1038/ng.823 PMID: 21552263; PubMed Central PMCID: PMCPMC4295936.

44. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics C, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015; 47(3):291–5. Epub 2015/02/03. https://doi.org/10.1038/ng.3211 PMID: 25642630; PubMed Central PMCID: PMCPMC4495769.

45. Eskin E. Discovering Genes Involved in Disease and the Mystery of Missing Heritability. Commun Acm. 2015; 58(10):80–7. https://doi.org/10.1145/2817827 PubMed PMID: WOS:000361909500025.

46. Vilhjalmsson BJ, Nordborg M. The nature of confounding in genome-wide association studies. Nat Rev Genet. 2013; 14(1):1–2. Epub 2012/11/21. https://doi.org/10.1038/nrg3382 PMID: 23165185.

47. Sul JH, Bilow M, Yang WY, Kostem E, Furlotte N, He D, et al. Accounting for Population Structure in Gene-by-Environment Interactions in Genome-Wide Association Studies Using Mixed Models. PLoS Genet. 2016; 12(3):e1005849. Epub 2016/03/05. https://doi.org/10.1371/journal.pgen.1005849 PMID: 26943367; PubMed Central PMCID: PMCPMC4778803.

48. Kostem E, Eskin E. Improving the accuracy and efficiency of partitioning heritability into the contributions of genomic regions. Am J Hum Genet. 2013; 92(4):558–64. Epub 2013/04/09. https://doi.org/10.1016/j.ajhg.2013.03.010 PMID: 23561845; PubMed Central PMCID: PMCPMC3617385.

49. Gusev A, Lee SH, Trynka G, Finucane H, Vilhjalmsson BJ, Xu H, et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am J Hum Genet. 2014; 95(5):535–52. Epub 2014/12/03. https://doi.org/10.1016/j.ajhg.2014.10.004 PMID: 25439723; PubMed Central PMCID: PMCPMC4225595.